DHBW
Duale Hochschule
Baden-Württemberg
Karlsruhe

Comparison of Selected Model Order Reduction methods

Studienarbeit (T3_3101)

im Rahmen der Prüfung zum

Bachelor of Science (B.Sc.)

des Studienganges Angewandte Informatik

an der Dualen Hochschule Baden-Württemberg Karlsruhe

von

Florian Braun

Januar 2018

-Sperrvermerk-

Abgabedatum:	22. Mai 2023
Bearbeitungszeitraum:	14.10.2022 - 22.05.2023
Matrikelnummer, Kurs:	7433149, TINF20B1
Gutachter der Dualen Hochschule:	Lutz Gröll

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich meine Studienarbeit (T3_3101) mit dem Thema:

Comparison of Selected Model Order Reduction methods

gemäß § 5 der "Studien- und Prüfungsordnung DHBW Technik" vom 29. September 2017 selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Karlsruhe, den April 19, 2023

Nachname, Vorname

Abstract

- Deutsch -

Dies ist der Beginn des Abstracts. Für die finale Bachelorarbeit musst du ein Abstract in deinem Dokument mit einbauen. So, schreibe es am besten jetzt in Deutsch und Englisch. Das Abstract ist eine kurze Zusammenfassung mit ca. 200 bis 250 Wörtern.

Versuche in das Abstract folgende Punkte aufzunehmen: Fragestellung der Arbeit, methodische Vorgehensweise oder die Hauptergebnisse deiner Arbeit.

Contents

List of abbreviations

DE	Differential Equation
DFT	Discrete Fourier Transform
FEM	Finite Element Method
FT	Fourier Transform
FFT	Fast Fourier Transform
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
LS	Least Squares
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation

List of Figures

List of Tables

source-codeverzeichnis

1. Fundamentals

1.1. Heat Equation

The conduction of heat within a medium can be described using the following partial differential equation (PDE):

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u \tag{1.1}$$

With u being a function of space and time and α being a positive constant. For this paper u will be defined in terms of one spatial dimension:

$$u := u(x,t) \tag{1.2}$$
$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \tag{1.3}$$
$$x \in \chi \subset \mathbb{R} \quad t \in \tau \subset \mathbb{R} \tag{1.4}$$
$$x_0 \leq x \leq x_n \quad t_0 \leq t \leq t_n \tag{1.5}$$

[1]
In order to not only model the conduction of heat within a medium but also a heating process, a new function $h : \chi \times \tau \to \mathbb{R}$ is introduced:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} + h(x,t) \tag{1.6}$$

For this paper it is assumed that the initial condition is known:

$$f : \chi \to \mathbb{R} \tag{1.7}$$
$$u(x,t_0) = f(x) \tag{1.8}$$

1.1.1. Solving the Heat Equation using Fourier Transform

Applying the Fourier transform (FT) w.r.t x to 1.6 yields the inhomogeneous ordinary differential equation (ODE):

$$\hat{u} = \mathfrak{F}(u) \quad \hat{h} = \mathfrak{F}(h) \tag{1.9}$$
$$\frac{d}{dt}\hat{u} = -\alpha\omega^2\hat{u} + \hat{h} \tag{1.10}$$

A solution to 1.10 is given by:

$$\hat{u} = \hat{u}_0 + \hat{u}_p \tag{1.11}$$

Where $\hat{u}_0$ is the homogeneous solution and $\hat{u}_p$ is the particular integral. In order to solve this ODE for the particular integral $\hat{h}$ has to be known [2]. The choice of h is, except for some restrictions, arbitrary. Therefore an approximate solution to 1.10 $\hat{u}_a$ is obtained by the forward euler scheme:

$$\frac{d}{dt}\hat{u} \approx \frac{\Delta\hat{u}}{\Delta t} \tag{1.12}$$

$$\hat{u}_{t+1} = \hat{u}_t + \Delta t(-\alpha\omega^2\hat{u} + \hat{h}) \tag{1.13}$$

$$\hat{u}_a = [\hat{u}_{t_0}, ..., \hat{u}_{t_n}] \tag{1.14}$$

In order to apply the euler scheme successfully an initial condition $\hat{u}_0$ has to be known. This initial condition is obtained by applying the discrete Fourier transform (DFT) to an initial temperature distribution along x:

$$\hat{u}_0 = \mathfrak{F}(f(x)) \tag{1.15}$$

[3]
The forward euler scheme is used here because it is fairly easy to implement. By applying the inverse discrete Fourier transform (IDFT) to $\hat{u}_a$ an approximate solution to 1.6 can be obtained. To decrease computing time, the fast Fourier transform (FFT) and inverse fast Fourier transform (IFFT) is used instead of the DFT and IDFT. A problem with solving the heat equation in this way is that it cannot handle boundary conditions. In order to solve the heat equation as initial boundary value problem (IBVP) the finite element method is used.

1.2. Finite Element Method

The finite element method (FEM) is a method to approximate solutions for differential equations (DE) within a certain domain Ω. This is done by discretizing the spatial domain. Assume that a DE is given by:

$$m, n \in \mathbb{N} \quad \zeta \in \Omega \subset \mathbb{R} \quad m \geq 1 \tag{1.16}$$

$$\frac{\partial^m y}{\partial \zeta^m} - g(y) = r(\zeta, t) \tag{1.17}$$

It is assumed that g is a linear function that can also contain partial derivatives of y w.r.t. time, y takes the value 0 at the boundary Γ and $y(\zeta, 0) = f(\zeta)$. An approximate solution to

y is given by μ, which is expressed as a sum of basis functions contained in the set ϕ:

$$\mu(\zeta, t) = \sum_{j=1}^{N} c_j(t)\phi_j(\zeta) \tag{1.18}$$

The residual is defined as:

$$\mathfrak{r} = \frac{\partial^m \mu}{\partial \zeta^m} - g(\mu) - r(\zeta, t) \tag{1.19}$$

Furthermore the residual is required to be orthogonal to all basis functions:

$$\langle \mathfrak{r}, \phi_k \rangle = 0 \quad \forall \phi_k \in \phi \tag{1.20}$$

Since the functions in ϕ are known, it is only required to find the coefficients $c_j(t)$ in 1.18. To find those coefficients 1.20 needs to be expressed as follows:

$$\int_\Omega \frac{\partial^m \mu}{\partial \zeta^m}\phi_k \, d\zeta - \int_\Omega g(\mu)\phi_k \, d\zeta = \int_\Omega r(\zeta, t)\phi_k \, d\zeta \quad \forall \phi_k \in \phi \tag{1.21}$$

If μ is substituted with 1.18 the following is obtained:

$$\sum_{j=1}^{N}((\int_\Omega \frac{\partial^m \phi_j}{\partial \zeta^m}\phi_k \, d\zeta)c_j(t) - g((\int_\Omega \phi_k \phi_j d\zeta)c_j(t))) = \int_\Omega r(\zeta, t)\phi_k \, d\zeta \quad \forall \phi_k \in \phi \tag{1.22}$$

It is also necessary to apply the divergence theorem to the first integral term taking into account that y at Γ is 0. Since ζ is one dimensional, the divergence theorem becomes integration by parts:

$$\int_\Omega \frac{\partial^m \phi_j}{\partial \zeta^m}\phi_k \, d\zeta = -\int_\Omega \frac{\partial^{m-1} \phi_j}{\partial \zeta^{m-1}}\frac{\partial \phi_k}{\partial \zeta} \, d\zeta \quad \forall \phi_k \in \phi \tag{1.23}$$

Combining 1.22 and 1.23 yields:

$$-\sum_{j=1}^{N}((\int_\Omega \frac{\partial^{m-1} \phi_j}{\partial \zeta^{m-1}}\frac{\partial \phi_k}{\partial \zeta} \, d\zeta)c_j(t) + g((\int_\Omega \phi_k \phi_j d\zeta)c_j(t))) = \int_\Omega r(\zeta, t)\phi_k \, d\zeta \quad \forall \phi_k \in \phi \tag{1.24}$$

This formulation leads to a system of ODEs or a system of linear equations that can be solved either analytically or numerically.

1.2.1. Solving the Heat Equation using FEM

This formulation of FEM can be applied to 1.6:

$$\Omega = \chi \quad \Gamma = \{x_0, x_n\} \tag{1.25}$$

$$y(\zeta, t) = -u(x, t) \quad g(u) = -\frac{1}{\alpha}\frac{\partial u}{\partial t} \tag{1.26}$$

$$m = 2 \quad r(\zeta, t) = \frac{1}{\alpha}h(x, t) \tag{1.27}$$

$$u(x, 0) = f(x) \quad u(x_0, t) = 0 \quad u(x_n, t) = 0 \tag{1.28}$$

The set of basis functions is defined as a set of piecewise linear functions with constant step size Δx:

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/\Delta x), & x_{j-1} \le x < x_j \\ (x_{j+1} - x)/\Delta x), & x_j \le x < x_{j+1} \\ 0, \quad \text{otherwise} \end{cases} \tag{1.29}$$

[4]
The step size Δx is defined by $\Delta x = \frac{x_n - x_0}{n-1}$. This results in the following system of ODEs:

$$\sum_{j=1}^{N}(\int_{\chi}\phi_j\phi_k dx)\frac{dc_j}{dt} = \alpha\sum_{j=1}^{N}(-\int_{\chi}\frac{d\phi_j}{dx}\frac{d\phi_j}{dx}dx)c_j(t) + \int_{\chi}h(x,t)\phi_k dx \quad \forall\phi_k \in \phi \tag{1.30}$$

Since $\phi_k\phi_j \neq 0$ for $k = j, k = j \pm 1$ (1.30) can be expressed as

$$q_1\dot{c}_{j-1} + q_2\dot{c}_j + q_1\dot{c}_{j+1} = p_1c_{j-1} + p_2c_j + p_1c_{j+1} + H(x,t) \tag{1.31}$$

where p and q are some constants derived later. For $j = 1$ and $j = n$ some new coefficients c_0, c_{n+1} are introduced. To force the boundary condition they are set to zero $c_0 = c_{n+1} = 0$. It follows that $h(x,t)$ and $c_j(0)$ have to satisfy the boundary condition too.
Using matrix notation this becomes:

$$M^{N\times N}, K^{N\times N} \tag{1.32}$$

$$M\dot{c} = Kc + d \tag{1.33}$$

The matrices M and K can be easily computed (Appendix A.1):

$$m_{ij} = \begin{cases} \frac{2\Delta x}{3}, & k = j \\ \frac{\Delta x}{6}, & |k-j| = 1 \\ 0, & otherwise \end{cases} \qquad k_{ij} = \begin{cases} \frac{-2\alpha}{\Delta x}, & k = j \\ \frac{\alpha}{\Delta x}, & |k-j| = 1 \\ 0, otherwise \end{cases} \tag{1.34}$$

Furthermore to solve this system of ODEs numerically an initial condition c_0 has to be known [3]. It can be obtained using a least squares (LS) approach:

$$\sum_{j=1}^{N} \langle \phi_j, \phi_k \rangle c_j(0) = \langle f, \phi_k \rangle \quad \forall \phi_k \in \phi \tag{1.35}$$

$$Mc_0 = F \tag{1.36}$$

$$c_0 = M^{-1}F \tag{1.37}$$

[5]. However since the basis functions are piecewise linear functions, the coefficients are $c_0 = \left[f(0), f(\Delta x), ..., f(n)\right]^T$ [**Gustafsson2011g**]. It is necessary to approximate d for each point in time using numerical integration schemes. By assuming that the following decomposition holds:

$$h(x,t) = \sum_{i=1}^{N} \phi_i(x) \upsilon(t) \tag{1.38}$$

the vector d in 1.33 can be expressed as $M\upsilon(t)$. Observe that by multiplying 1.33 with M^{-1} (A.2) yields a system of ODEs:

$$\dot{c} = M^{-1}Kc + M^{-1}M\upsilon \tag{1.39}$$

This system of ODEs can be solved using an euler scheme:

$$c_{t+1} = \Delta t(M^{-1}Kc_t + I\upsilon) + c_t \tag{1.40}$$

Using 1.18 and the computed coefficients c the function $u(x,t)$ can be approximated. However this is equivalent to linear interpolation between c_n and c_{n+1} (Appendix A.3).

1.3. Singular Value Decomposition

The Singular Value Decomposition (SVD) is a matrix factorization with guaranteed existence. It can be used to obtain low rank approximations of a matrix or pseudo inverses for ill posed linear system of equations. It is also related to FT by providing a data specific set of orthogonal bases instead of a generic set of sines and cosines. For this paper the SVD will be used for generating low rank approximations of matrices [6].

1.3.1. Properties

A matrix $X \in \mathbb{C}^{n \times m}$ can be decomposed in the following way:

$$X = U\Sigma V^* \tag{1.41}$$

Here $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{m\times m}$ are unitary matrices and $\Sigma \in \mathbb{R}^{n\times m}$ is a real valued ordered diagonal matrix. The columns of U provide a set of orthonormal basis vectors for the column space of X, V contains orthonormal basis vectors for the row space of X. The matrix Σ assigns a magnitude ('importance') to the product of U and V^* [7]. Since U and V are unitary they have the following property:

$$U^*U = UU^* = I \tag{1.42}$$
$$V^*V = VV^* = I \tag{1.43}$$

[8]
In case $n \geq m$ the so called economy SVD can be used to factorize the matrix X:

$$X = \begin{bmatrix}\hat{U} & \hat{U}^\perp\end{bmatrix}\begin{bmatrix}\hat{\Sigma}\\ 0\end{bmatrix}V^* = \hat{U}\hat{\Sigma}V^* \tag{1.44}$$

The economy SVD omits rows only containing zeros in Σ and the according columns of U. Therefore the dimensionality of $\hat{U}$ and $\hat{\Sigma}$ is less or equal to the dimensionality of U and Σ [6].

1.3.2. Hierarchy of correlations

As already stated, the matrix Σ assigns a magnitude to UV^*. This magnitude can be seen as the square of the variance the bases in U and V capture. Assume that X^*X and XX^* denote correlation matrices [6]. A correlation matrix is a matrix that stores correlation coefficients between multiple measurements. If X has the following properties X^*X and XX^* are correlation matrices:

1.) The column vectors of X have to be zero mean:

$$X = \begin{bmatrix}x_1, \cdots, x_m\end{bmatrix} \tag{1.45}$$
$$\frac{1}{n}\sum_{j=1}^{n} x_{ij} = \mu_i = 0 \quad \forall 1 \leq i \leq m \tag{1.46}$$

2.) The column vectors of X have to be normalized:

$$(\sum_{j=1}^{n} x_{ij}^2)^{\frac{1}{2}} = \bar{x_i} = 1 \quad \forall 1 \leq i \leq m \tag{1.47}$$

The correlation coefficient between two column vectors of X is calculated as follows:

$$corr(x_i, x_{i'})\frac{(\sum_{j=1}^{n} x_{ij} - \bar{x_i})(\sum_{j=1}^{n} x_{i'j} - \bar{x_{i'}})}{(\sum_{j=1}^{n}(x_{ij} - \bar{x_i})^2)^{\frac{1}{2}}(\sum_{j=1}^{n}(x_{i'j} - \bar{x_{i'}})^2)^{\frac{1}{2}}} \tag{1.48}$$

[9]
Since all column vectors have zero mean and are normalized this becomes:

$$corr(x_i, x_{i'}) = cov(x_i, x_{i'}) = x_i^* x_{i'} \tag{1.49}$$

[10]
This resembles the entries of XX^* and X^*X. A vector e that maximizes the variance of the projection of x_i onto e with the restriction $\|e\| = 1$ are the eigenvectors of the according correlation matrix. The eigenvalue of e denoted as λ is equivalent to the variance of x_i projected onto e [11].
Since X can be de-constructed using the SVD, XX^* and X^*X are equal to:

$$XX^* = U \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix} V^* V \begin{bmatrix} \hat{\Sigma} & 0 \end{bmatrix} U^* = U \begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} U^* \tag{1.50}$$

$$X^*X = V \begin{bmatrix} \hat{\Sigma} & 0 \end{bmatrix} U^* U \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix} V^* = V \hat{\Sigma}^2 V^* \tag{1.51}$$

By multiplying U and V respectively on the right side 1.50 and 1.51 become:

$$XX^*U = U \begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} \tag{1.52}$$

$$X^*XV = V \hat{\Sigma}^2 \tag{1.53}$$

This shows that V contains the eigenvectors of the row-wise correlation matrix and U contains the eigenvectors of the column-wise correlation matrix. The matrix Σ contains the roots of the according eigenvalues and is thereby related to the variance. By ordering U, Σ and V by the entries of Σ in a descending order he first row of U and V contain the most important basis vectors [6].

1.3.3. Low-rank approximation

A useful property of the SVD is that it can be used to find an hierarchy of rank-r approximation for a given matrix X. An matrix $\tilde{X}$ that approximates X is obtained by:

$$\tilde{X} = arg\,min \|X - \tilde{X}\|_F = \tilde{U} \tilde{\Sigma} \tilde{V}^* \tag{1.54}$$

$$s.t. rank(\tilde{X}) = r \tag{1.55}$$

Here $\tilde{U}$ and $\tilde{V}$ denote matrices obtained taking the first r columns of U and V. The matrix $\tilde{\Sigma}$ is a $r \times r$ sub-block of Σ. This is also known as the Eckard-Young theorem. The variance

captured by $\tilde{X}$ can be calculated in the following way:

$$cumvar_r(\Sigma) = \frac{\sum_{i=1}^{r} \sigma_i}{trace(\Sigma)} \tag{1.56}$$

$$var(\Sigma) = \frac{diag(\Sigma)}{trace(\Sigma)} \tag{1.57}$$

[6]

1.3.4. Example low-rank approximation

As an example suppose there is a matrix $X \in \mathbb{Z}^{4\times4}$:

$$X = \begin{bmatrix} 3 & 1 & 5 & 5 \\ 4 & -4 & 5 & 0 \\ -4 & -2 & -4 & 3 \\ 5 & 1 & 5 & -4 \end{bmatrix} \tag{1.58}$$

By computing the SVD the matrices U, Σ and V are obtained. Now Σ can be used to calculate the cumulative variance for an rank-r approximation and the variance captured by each basis vector of U and V.

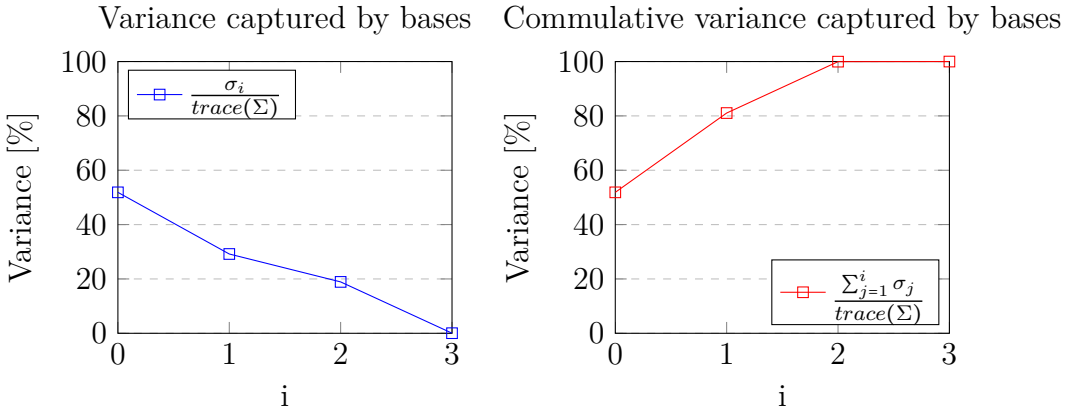

Figure 1.1.: Variance and cumulative variance captured by each column vectors of U and V

On 1.3.4 the commutative variance and the variance of each basis vector is plotted. Here X can be approximated using the first three leading basis vectors. This approximation captures already more than 99% of the variance. The resulting matrix $\tilde{X}_3$ looks as follows:

$$\tilde{X}_3 = \begin{bmatrix} 4.73 & -0.76 & 5.62 \\ 1.31 & -1.37 & 1.62 \\ -5.10 & -0.60 & 2.34 \end{bmatrix} \tag{1.59}$$

1.4. Fundamentals of Control Theory

1.4.1. Problems in Control Theory

There are mainly three different problems in control theory. To analyse a system it has to be identified first. A mathematical model has to be found that relates the input of that system to it's output. After a model has been found, it is helpful to simulate the system to find out which input results in which output. Simulations have to fit the system and offer the advantage over real experiments of reducing cost and risks attached to them. The last of the three major problems is the control problem. Now that the underlying model is known and the system can be simulated, a controller for that system can be designed. Here the controller is a system that feeds some input to the system that is to be controlled to generate some desired output [12]. For this paper the second problem is the most important one, since the goal of this paper is to compare some methods to make simulations of a system more efficient.

1.4.2. LTI Systems

A linear and time invariant (LTI) system is a kind of system that fulfils the following conditions:

Homogenity: For a system to be a LTI system, the condition that the output of a system $y(t)$ relates to the input $u(t)$ by a linear operator f:

$$y(t) = f(u(t)) \tag{1.60}$$
$$cy(t) = f(cu(t)) \tag{1.61}$$

This means that if the input u changes in magnitude by a constant factor of c the output y also changes by the factor of c.

Superposition: The second requirement is superposition:

$$y_1(t) + y_2(t) = f(u_1(t) + u_2(t)) \tag{1.62}$$

The sum of two outputs y_1 and y_2 has to equal the output of the system with the sum of the inputs u_1 and u_2 as input.

Time Invariance: A system that is time invariant has the property that if the input u is shifted in time by some constant τ the resulting output is also shifted in time by the same constant:

$$y(t-\tau) = f(u(t-\tau)) \tag{1.63}$$

LTI systems are important because they can be used to approximate non LTI systems over some region. This is useful since LTI systems are well understood [13].

1.4.3. State Space Representation

A LTI system can be expressed as a system of ODEs that relates the input u to its output y. The vector x denotes the states of a system.

$$\dot{x} = Ax + Bu \quad x(t_0) = x_0 \tag{1.64}$$

$$y = Cx + Du \tag{1.65}$$

The matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{q \times n}$ and $D \in \mathbb{R}^{q \times m}$ are constant matrices. [14]

1.4.4. State Space of Discrete Time Systems

A system can represented as discrete system in the following form:

$$x_{k+1} = A_d x_k + B_d x_k \tag{1.66}$$

$$y_{k+1} = C_d x_x + D_d u_k \tag{1.67}$$

$$x_k = x_{k\Delta t} \tag{1.68}$$

The matrices A_d to D_d can be obtained from the continuous system in the following way:

$$A_d = e^{A\Delta t} \quad B_d = \int_0^{\Delta t} e^{A\tau} B d\tau \tag{1.69}$$

$$C_d = C \quad D_d = D \tag{1.70}$$

[15]

1.4.5. Transfer Function

To understand the transfer function the impulse response has to be defined first. The impulse response of a system is given by choosing the following function as input to the system:

$$g(t) = f(\delta(t)) \tag{1.71}$$

The function $\delta(t)$ is the delta function:

$$\delta(t) = \begin{cases} \infty, & t = 0 \\ 0 & else \end{cases} \tag{1.72}$$

It has the property that the integral of this function is 1:

$$\int_{t=-\infty}^{\infty} \delta(t)dt = 1 \tag{1.73}$$

If the $g(t)$ is convoluted with a function $v(t)$ the response of the system for $v(t)$ as input is obtained:

$$v(t) * g(t) = \int_{\tau=0}^{\infty} v(\tau)g(t-\tau)d\tau = f(v(t)) \tag{1.74}$$

[16]
By using the convolution theorem:

$$\mathfrak{L}\{g(t)\}\mathfrak{L}\{v(t)\} = \mathfrak{L}\{g * v\}(t) \tag{1.75}$$

[17]
the transfer function is defined as:

$$\mathfrak{L}\{f(\delta(t))\} \tag{1.76}$$

This definition is useful since the response of a system for a given input $u(t)$ can be determined by the result of the multiplication $G(s)U(s)$ [16]. Alternatively if a system is given in state space representation the transfer function can be expressed as:

$$G(s) = C(sE - A)^{-1}B + D \tag{1.77}$$

[14]

1.4.6. Controallability and Observability

For LTI systems in state space representation it can be determined which states are to what degree controllable and observable. This can be computed by the controllability and observability gramians:

$$W_c = \lim_{t\to\infty} \int_0^t e^{A\tau}BB^*e^{A^*\tau}d\tau \tag{1.78}$$

$$W_o = \lim_{t\to\infty} \int_0^t e^{A^*\tau}C^*Ce^{A\tau}d\tau \tag{1.79}$$

The degree of controllability for a state x can be determined by x^*W_cx. If the result of this calculation is large, the system is controllable in the x direction. By swapping W_c with W_o the degree of observability in state x can be computed: x^*W_ox. Again if the resulting value is large, the system can be observed well in state x [18].

2. Model Order Reduction

2.1. Introduction

Model Order Reduction (MOR) is a technique to reduce the computational effort of simulating a system using mathematical models. This is done by using two different approaches. The first one is using a numerical approach and the second approach is based on system theory. The first method discussed is called Proper Orthogonal Decomposition. After that Balanced and Modal Truncation will be discussed. The last method discussed will be Hankel-norm approximation [19].

2.2. Proper Orthogonal Decomposition

The proper orthogonal decomposition (POD) is a method for model order reduction. The reduction in computational effort is done by approximating a solution to a PDE using an orthogonal expansion. The basis functions are obtained by decomposing a set of solutions for the PDE using the SVD.

2.2.1. Orthogonal Expansion

A function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ can be represented by the series:

$$f(x,t) = \sum_{i=1}^{\infty} a_i(t)\phi_i(x) \tag{2.1}$$

$$x, t \in \mathbb{R} \tag{2.2}$$

Here all $\phi(x)$ are orthogonal basis functions.

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{else} \end{cases} \tag{2.3}$$

By using a finite sum instead of the entire series f can be approximated:

$$\tilde{f}(x,t) = \sum_{i=1}^{n} a_i(t)\phi_i(x) \tag{2.4}$$

Since all ϕ are known, 2.4 has to be solved for the set $\{a_0, ..., a_n\}$. In case f is known, the solutions contained in this set can be calculated as:

$$a_i(t) = \frac{\langle \phi_i, f \rangle}{\langle \phi_i, \phi_i \rangle} \tag{2.5}$$

[20]

2.2.2. Proper Orthogonal Decomposition for PDEs

Since a PDEs are equations in terms of partial derivatives, the notation $\mathscr{P}(\partial x)u$ is introduced, which denotes a differential operator in terms of the spatial variables x for a function $u(x, t, p$. The vector p contains some parameters. This is done to provide a more abstract way to denote PDEs:

$$\frac{\partial u}{\partial t} = \mathscr{P}(\partial x)u \tag{2.6}$$

[21]
Solving for u is often difficult to impossible. A method that is often used to solve PDEs is called separation of variables. This separation of variables assumes, that the underlying solution $u(x, t)$ can be expressed by 2.2, to solve for $a_k, 0 \leq k \leq n$. Since it is not practical to compute an infinite series, u only gets approximated by using 2.4 instead:

$$\sum_{i=1}^{n} \frac{\partial a_i}{\partial t}\phi_i = \sum_{i=1}^{n} \mathscr{P}(\partial x)\phi_i a_i \tag{2.7}$$

By discretizing the spatial dimension got along x 2.2.4 can be expressed using matrix notation:

$$\Phi = \left[\phi_0, ..., \phi_n\right] \tag{2.8}$$

$$\Phi \frac{d}{dt}a = \mathscr{P}(\partial x)\Phi a \tag{2.9}$$

Since the solution u is unknown 2.5 cannot be computed. However, the fact, that all ϕ are orthogonal to each other, can be used to solve for all a_k. It can be done by computing the inner product of 2.2.4 with all basis functions:

$$\langle \sum_{i=1}^{n} \frac{\partial a_i}{\partial t}\phi_i, \phi_k \rangle = \langle \sum_{i=1}^{n} \mathscr{P}(\partial x)\phi_i a_i, \phi_k \rangle \quad \forall 0 \leq k \leq n \tag{2.10}$$

This resembles the Galerking projection. In matrix notation it can be expressed:

$$\Phi^* \Phi \frac{d}{dt}a = \Phi^* \mathscr{P}(\partial x)\Phi a \tag{2.11}$$

By considering 2.3 the equation 2.10 can be expressed as a system of ODEs which can be solved:

$$\frac{d}{dt}a = \Phi^* \mathscr{P}(\partial x)\Phi a \tag{2.12}$$

After the vector of coefficients a for each time step has been computed, the solution can be assembled:

$$u(x,t) \approx \Phi(x)a(t) \tag{2.13}$$

[22]

2.2.3. Selection of Basis Vectors

As discussed in the previous section, a set of basis vectors can be used to generate approximate solutions to PDEs. However it was not discussed how those basis functions are chosen. For POD to work, a so-called snapshot matrix X has to be available. This snapshot matrix stores a set of solutions where x_k is the solution for a PDE at time step $k\Delta t$:

$$X = \left[x_1, ..., x_m\right] \tag{2.14}$$

The solutions can be obtained by conducting an experiment on a physical system that is described by the PDE or by simulating the evolution of that PDE. In this paper the solution is obtained using FEM 1.2. SVD is used to decompose the snapshot matrix X. Since the columns of the snapshot matrix contain spatial information at a given point in time and the basis vectors are supposed to encode the spatial information of a solution u the r most dominant left singular values have to be extracted:

$$\tilde{X} = \tilde{U}\tilde{\Sigma}\tilde{V}^* \tag{2.15}$$

$$\Phi = [u_1, ..., u_r] \tag{2.16}$$

Here the FEM solution is obtained by discretizing the PDE into a large number(n) of spatial nodes. This results in a high dimensional system of ODEs. Since Φ contains only r column ($r << n$) vectors a reduction in order can be achieved. The number of modes r is determined by the variance that has to be persevered. Figure 2.1 shows the variance each mode stores. Here the first two modes alread store almost 97% of the variance.

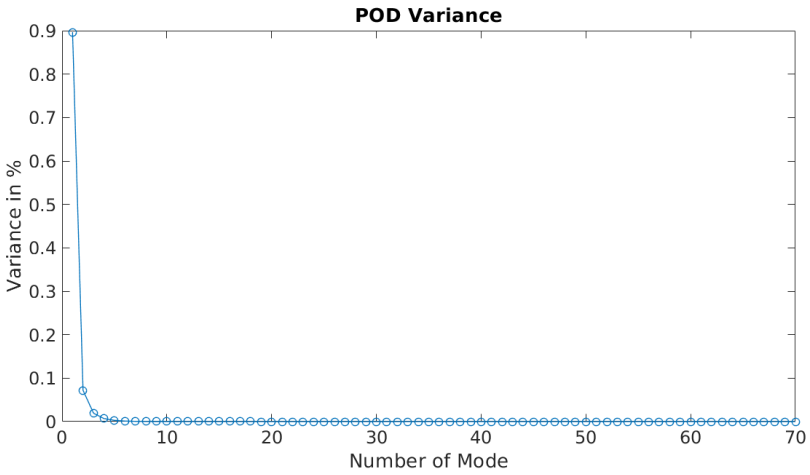

Figure 2.1.: Variance of Modes

These two modes are displayed on 2.2.

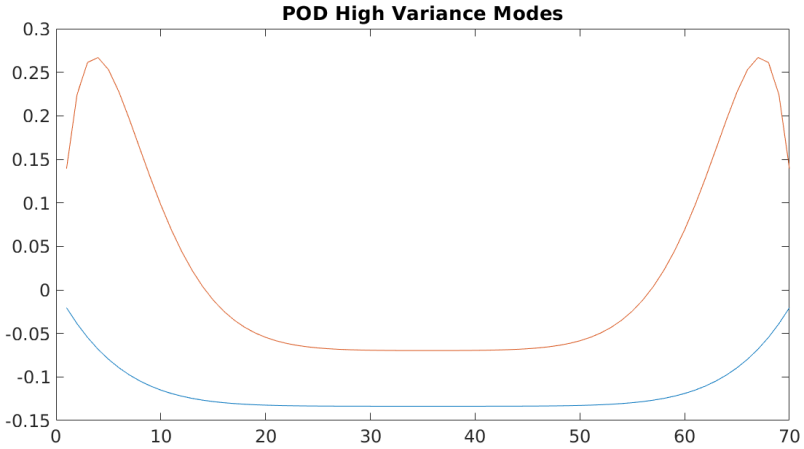

Figure 2.2.: Leading High Variance Modes

Setting the variance too high inlcudes also low variance modes that will introduce some errors. Some low variance modes are displayed on 2.3.

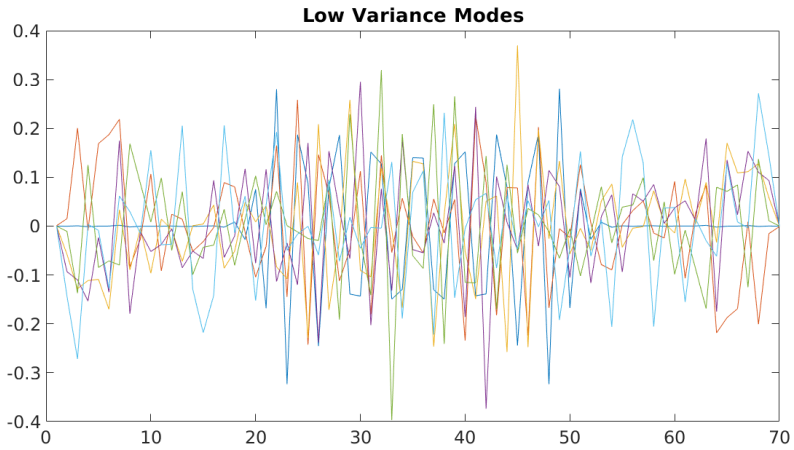

Figure 2.3.: Modes 30 to 35

Figure 2.1 shows that modes 30 to 35 almost have zero contribution to the snapshot matrix. Therefore including them will introduce noise to the approximation and will increase computations. However the benefits of this reduced order model are only relevant after the PDE has been solved once using a high dimensional system of ODEs [22].

2.2.4. POD for Heat Equation

In order to apply pod to heat equation a snapshot matrix X has to be generated. This is done by the FEM solver described in 1.2. After that X gets decomposed using the SVD. The modes contained in Φ is obtained by truncating the left singular vectors of X according to 2.16. Substituting $\mathscr{P}$ in with heat equation 1.6 results in the following:

$$\Phi \frac{\partial a}{\partial t} = \alpha \frac{\partial^2 \Phi}{\partial x^2} a + h \tag{2.17}$$

$$\frac{\partial a}{\partial t} = \alpha \Phi^* \frac{\partial^2 \Phi}{\partial x^2} a + \Phi^* h \tag{2.18}$$

This system of ODEs can now be solved using a Runge-Kutta scheme. Note that Φ contains numeric values. Therefore derivatives are unstable, especially at the first and last entries of the column vectors of Φ. To reduce this problem the method used to compute the second derivative of Φ is the so-called spectral derivative. The discrete spectral derivative works by computing the FFT of a vector. That vector is multiplied by $(ik)^d$, d is the order of the derivative and k are the discrete wave numbers. After that step the IFFT is applied to obtain

the derivative of the original vector.

$$f \in \mathbb{C}^n, \quad k = [-\frac{n}{2} \cdots \frac{n}{2}]^T \tag{2.19}$$

$$\frac{df}{dx} = \mathfrak{F}^{-1}\{i\frac{2\pi k}{n}\mathfrak{F}\{f\}\} \tag{2.20}$$

$$\frac{d^2f}{dx^2} = \mathfrak{F}^{-1}\{-\frac{2\pi k}{n}\mathfrak{F}\{f\}\} \tag{2.21}$$

$$\tag{2.22}$$

[23]
Figure 2.4 shows a solution to heat equation using FEM and an approximation using POD with $u(x,t) = 0, x_0 = 10000^{70\times1}, n = 70, variance = 90\%$. It is clearly visible that there are some differences, especially near the edges. However a more detailed analysis of the results are covered in chapter 4.

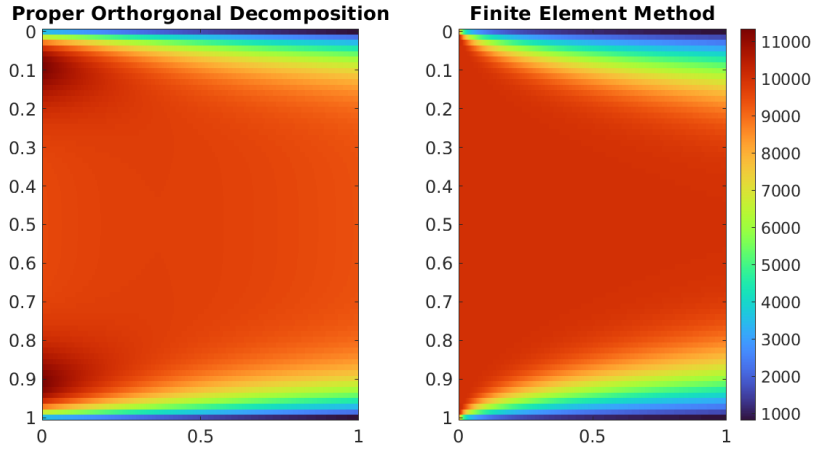

Figure 2.4.: FEM solution and POD approximation for heat equation

2.3. Balanced Truncation

Balanced truncation is a method for model order reduction. The goal of balanced truncation is to approximate a system using only the most relevant modes of the system. The difference to POD is that the modes are not selected by the variance they capture but by the controllability and observability of the modes. This is done by finding a coordinate transform.

2.3.1. Balanceing Coordinate Transform

To find a reduced order model using balanced truncation a orthonormal coordinate transform is applied: $x = Tz$. This yields a new system:

$$\dot{z} = \hat{A}z + \hat{B}u \tag{2.23}$$

$$y = \hat{C}z + Du \tag{2.24}$$

$$\hat{A} = T^{-1}AT \quad \hat{B} = T^{-1}B \quad \hat{C} = CT \tag{2.25}$$

The gramians of this ROM can be obtained by applying 1.79 and 1.78 to 2.25. This yields $\hat{W}_c = T^{-1}W_cT^{-*}$ and $\hat{W}_o = T^*W_oT$ with $T^{-*} := (T^{-1})^* := (T^*)^{-1}$. A requirement T has to satisfy is that it has to make the observability and controllability gramians of the ROM equal and diagonal.

$$\hat{W}_c = \hat{W}_o = \Delta \tag{2.26}$$

$$\hat{W}_c\hat{W}_o = \Delta^2 \tag{2.27}$$

$$T^{-1}W_cW_oT = \Delta^2 \tag{2.28}$$

$$W_cW_oT = T\Delta^2 \tag{2.29}$$

Since Δ is a diagonal matrix, 2.29 is equal to the eigendecomposition of W_cW_o. Therefore T contains the eigenvectors of W_cW_o. The values contained in Δ are known as hankel singular values. However T needs to be rescaled to make $\hat{W}_c$ and $\hat{W}_o$ equal. Here T_u denotes the unscaled eigenvectors of this eigendecomposition that yields gramians that are not equal to each other:

$$T_u^{-1}W_cT_u^{-*} = \Delta_c \tag{2.30}$$

$$T_u^*W_cT_u = \Delta_o \tag{2.31}$$

Scaling T_u by some diagonal matrix Δ_s results in $\Delta_c = \Delta_o$:

$$\Delta_s = \Delta_c^{\frac{1}{4}}\Delta_o^{-\frac{1}{4}} \tag{2.32}$$

$$T = T_u\Delta_s \tag{2.33}$$

Another important property of this transform is that the new coordinates are hierarchically ordered by observability and controllability. It can be shown by deriving some unit vector ζ that maximizes the controllability and observability:

$$\zeta = arg\max \zeta^*W_cW_o\zeta \quad s.t. \|\zeta\|_2^2 = 1 \tag{2.34}$$

$$\frac{d}{d\zeta}\zeta^*W_cW_o\zeta - 2\lambda\zeta = 0 \tag{2.35}$$

As shown here [24] the remaining derivative can be solved in the following way:

$$\frac{d}{dx}x^*Ax = 2Ax \tag{2.36}$$

This holds if A is symmetric. Here both W_c and W_o share the same set of eigenvectors A.15 and 2.31, therefore they commute [25]. This means that the product of W_c and W_o is also symmetric [26]. Applying 2.36 to 2.35 yields:

$$W_cW_o\zeta = \lambda\zeta \tag{2.37}$$

Since λ is the Lagrange multiplier, it is a scalar. It is clear that ζ is a eigenvector. Since the eigenvalues in Δ contain information about how much each eigenvector gets scaled by multiplying it with W_cW_o the eigenvectors can be ordered by controllability and observability.

2.3.2. Mode Truncation

Since the goal of balanced truncation is to find a ROM of rank r that approximates the original system of rank n with $r << n$ it is necessary to truncate the balanced system. This yields the following system:

$$\frac{d\tilde{x}}{dt} = \tilde{A}\tilde{X} + \tilde{B}u \tag{2.38}$$

$$y = \tilde{C}\tilde{x} + \tilde{D}u \tag{2.39}$$

The new state vector $\tilde{x}$ is defined as:

$$\tilde{x} = \begin{bmatrix} z_1 \\ \vdots \\ z_r \end{bmatrix} \quad \tilde{z} = \begin{bmatrix} z_{r+1} \\ \vdots \\ z_n \end{bmatrix} \quad z = \begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} \tag{2.40}$$

$$T = \begin{bmatrix} \Psi & T_t \end{bmatrix} \quad T^{-1} = S = \begin{bmatrix} \Phi^* \\ S_t \end{bmatrix} \tag{2.41}$$

By substituting 2.40 and 2.41 into the system in 2.23 and 2.24 the system becomes:

$$\frac{d}{dt}\begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} = \begin{bmatrix} \Phi^*A\Psi & \Phi^*AT_t \\ S_tA\Psi & S_tAT_t \end{bmatrix}\begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} + \begin{bmatrix} \Phi^*B \\ S_tB \end{bmatrix}u \tag{2.42}$$

$$y = \begin{bmatrix} C\Psi & CT_t \end{bmatrix}\begin{bmatrix} \tilde{x} \\ \tilde{z} \end{bmatrix} + Du \tag{2.43}$$

However the only relevant part of this system is:

$$\frac{d\tilde{x}}{dt} = \Phi^* A \Psi \tilde{x} + \Phi^* B u \tag{2.44}$$

$$y = C \Psi \tilde{x} + D u \tag{2.45}$$

since this is the only part necessary to calculate $\tilde{x}$ [27].

2.3.3. Computing Balanced Truncation

Since the gramians for controllability and observability are too expensive to compute for large systems the so called empirical gramians are used as an approximation. The empirical gramians are calculated by using the discrete-time system matrices from 1.66 and 1.67:

$$\mathscr{C}_d = [B_d A_d B_d \cdots d A^{m_o - 1} B_d] \tag{2.46}$$

$$\mathscr{O}_d = \begin{bmatrix} C_d \\ C_d A_d \\ C_d A_d^{m_o - 1} \end{bmatrix} \tag{2.47}$$

$$W_c^e = \mathscr{C}_d^* \mathscr{C}_d \tag{2.48}$$

$$W_o^e = \mathscr{O}_d^* \mathscr{O}_d \tag{2.49}$$

$$m_o, m_c << rank(A) \tag{2.50}$$

Now these empirical gramians can be used for obtaining the balancing coordinate transform [27].

2.3.4. State Space Representation from Heat Equation

To apply balanced truncation to the heat equation 1.6 a state space representation of that system has to be found first. Note that the system of ODEs resulting from FEM 1.39 resembles a state space representation:

$$x := c \quad x_0 := c_0 \tag{2.51}$$

$$A := M^{-1} K \quad B := I \tag{2.52}$$

Since the system is realized as simulation all states can be accurately measured $C := I$ and there is no feed through $D := 0_{qm}$. Now the already described steps for balanced truncation can be applied to the system.
Figure 2.5 shows a solution to heat equation using FEM and an approximation using BT with $u(x,t) = 0, x_0 = 10000^{70 \times 1}, n = 70, n_{approx} = 10$.

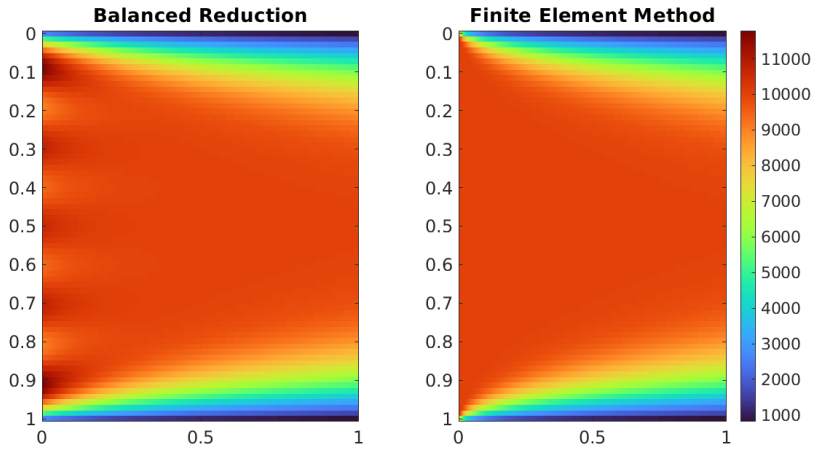

Figure 2.5.: FEM solution and BT approximation for heat equation

2.4. Modal truncation

2.4.1. Diagonal Canonical Form

A LTI system (A, B, C, D) with transfer function $G(s)$ can be written in the following way:

$$I = \{1, 2, \cdots, n\} \tag{2.53}$$

$$G(s) = D + \sum_{i \epsilon I} \frac{\Phi_i}{s - \lambda_i} \tag{2.54}$$

The set $\Lambda = \{\lambda_1, \cdots, \lambda_n\}$ denotes the eigenvalues of the matrix A in state space. This resembles the partial fraction decomposition of $G(s)$ [28]. The residue Φ can be computed using the eigendecomposition of A and using the eigenvectors for a coordinate transform:

$$AX = X\Delta \quad X\zeta = x \tag{2.55}$$

The system matrices have to be transformed accordingly:

$$\hat{A} = X^{-1}AX = \Delta \tag{2.56}$$

$$\hat{B} = X^{-1}B \quad \hat{C} = CX \tag{2.57}$$

$$\hat{D} = D \tag{2.58}$$

Those transformed matrices are used to calculate the transfer function 1.77:

$$G(s) = CX(sI - \Delta)^{-1}X^{-1}B + D \tag{2.59}$$

$$= D + \begin{bmatrix} Cx_1 & \cdots & Cx_n \end{bmatrix} diag\{\frac{1}{s-\lambda_1}, \cdots, \frac{1}{s-\lambda_n}\} \begin{bmatrix} x_1^{-1}B \\ \vdots \\ x_n^{-1}B \end{bmatrix} \tag{2.60}$$

$$= D + \sum_{i \in I} \frac{c_i b_i^T}{s-\lambda_i} \Rightarrow \Phi_i = c_i b_i^T \tag{2.61}$$

[29]

2.4.2. Optimal Modal Truncation

The goal of modal truncation is to find a subset of the indices I such that only r elements are contained in this subset:

$$I_r \subseteq I, \quad |I_r| = r \tag{2.62}$$

Using this subset as indices in 2.54 a truncation is obtained, yielding a system defined by a new transfer function $\hat{G}(s)$. The set I_r has to be chosen such that the error $\|G(s) - \hat{G(s)}\|_{H_n}$ becomes minimal. Here H_n denotes the H_2 norm. As shown in [28] other norms are also usable but for simplicity only the stated norm is used. Furthermore in case some $\lambda_i \in \Lambda, Im(\lambda_i) > 0$ is a complex number $i \in I_r$, the index of the according complex conjugate eigenvalue has to selected too, if $\lambda_j = \bar{\lambda}_i \in \Lambda \wedge i \in I_r \Rightarrow j \in I_r$. This yields the following optimization problem:

$$G_\alpha(s) = D_\alpha + \sum_{i \in I} \alpha_i \frac{\Phi_i}{s-\lambda_i} \tag{2.63}$$

$$\min_\alpha \|G(s) - G_\alpha(s)\| \tag{2.64}$$

$$s.t. \quad \alpha^T \alpha = r \tag{2.65}$$

$$J = \{(\lambda_i, \lambda_j) \in I | \lambda_i \in \mathbb{C}, Im(\lambda_i) > 0, \lambda_i = \bar{\lambda_j}\} \tag{2.66}$$

$$a_{.,j} = -a_{.,l} = 1, \quad (i,j) \in J \tag{2.67}$$

$$A\alpha = 0 \tag{2.68}$$

Where $\alpha \in \{0,1\}^n$ is some binary vector with $\alpha^T\alpha = r$ that represents the selection of indices. By defining the error of that system as:

$$\epsilon_\alpha(s) = G(s) - G_\alpha \tag{2.69}$$

$$\|\epsilon_\alpha(s)\|_2^2 = \sum_{i,k\in I}(1-\alpha_i)\frac{tr(\phi_i\phi_k^T)}{-\lambda_i-\lambda_k}(1-\alpha_k) \tag{2.70}$$

$$\|\epsilon_\alpha(s)\|_2^2 = (1-\alpha)Q(1-\alpha) \tag{2.71}$$

$$q_{ij} = \frac{tr(\phi_i\phi_k^T)}{-\lambda_i-\lambda_k^*} \tag{2.72}$$

This leads to a new optimization problem:

$$\min_\alpha \quad (1-\alpha)Q(1-\alpha) \tag{2.73}$$

$$s.t. \quad \alpha^T\alpha = r \tag{2.74}$$

$$A\alpha = 0 \tag{2.75}$$

[28]

2.4.3. Applying Modal Truncation to Heat Equation

As described in 2.3.4 the system matrices (A, B, C, D) can be obtained for the given heat equation. Now by applying 1.77 yields the corresponding transfer function $G(s)$. By calculating the partial fraction decomposition as described in 2.54 the DCNF can be obtained. Since B and C are given by some identity matrix and D is a zero matrix the upper error bound of the error in H_2 norm is given as

$$\|\epsilon(s)\|_{H_2} \leq \sum_{i\in I_r^c}\frac{1}{\sqrt{-2Re(\lambda_i)}} \tag{2.76}$$

[28]. Since for stable systems $Re(\lambda) < 0 \forall \lambda \in \Lambda$ this upper bound can be expressed as follows with $\Lambda_\epsilon = \{Re(\lambda_i)|\lambda_i \in \Lambda i \in I_r^c\}$

$$\|\epsilon(s)\|_{H_2} \leq \sum_{i\in I_r^c}\frac{1}{\sqrt{2|Re(\lambda_i)|}} \leq \frac{|I_r^c|}{\sqrt{2|\max(\Lambda_\epsilon)|}} \tag{2.77}$$

$$\tag{2.78}$$

To minimize this upper bound $|\max(\Lambda_\epsilon)|$ has to be maximized. This is achieved by selecting the entries of Λ_ϵ such that it contains $n-r$ largest real parts of the eigenvalues in Λ.
Figure 2.6 shows a solution to heat equation using FEM and an approximation using MT with $u(x,t) = 0, x_0 = 10000^{70\times1}, n = 70, n_{approx} = 10$.

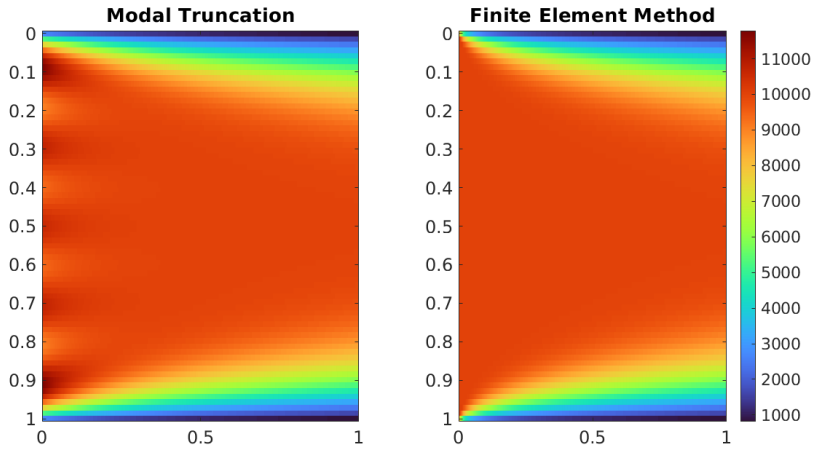

Figure 2.6.: FEM solution and MT approximation for heat equation

2.5. Hankel Norm Approximation

The goal of the Hankel Norm Approximation (HNA) is to find a reduced order system that approximates a system G such that the error in the so called Hankel norm becomes minimal. The Hankel norm is defined to be the largest hankel singular value

$$\|G(s)\|_H = \sigma_{max} = \sqrt{\lambda_{max}(W_c W_o)} \tag{2.79}$$

[30]. Therefore the problem can be stated as follows

$$G_r = arg\min \|G - G_r\|_H \tag{2.80}$$

2.5.1. System Spaces

For the following section it is necessary to define the four following system spaces: L_∞, H_∞, H_∞^- and $H_\infty^-(r)$.

L_∞ $G \in L_\infty$ iff $\sup_\omega \|G(i\omega)\| < \infty$.

H_∞ $G \in H_\infty$ iff $\forall \lambda \in \mathbb{C}_-$. Here λ are the eigenvalues of A.

H_∞^- $G \in H_\infty^-$ iff $G(-s) \in H_\infty$.

$H_\infty^-(r)$ $G \in H_\infty^-(r)$ iff $G \in L_\infty$ and $\lambda = \lambda_+ \cap \lambda_- = \{\lambda_1, ..., \lambda_n\}$ with $\lambda_\circ = \{\lambda_i \in \lambda | \lambda_i \in \mathbb{C}\_\circ\}$ and $|\lambda_-| \leq r$.

2.5.2. Optimal Solution

A lower bound for $\min \|G - G_r\|_H$ is established by lemma 7.1 in [31]

$$\min \|G - G_r\|_H \geq \sigma_{r+1} \tag{2.81}$$

.
Here σ_{r+1} is the $r+1$th largest hankel singular value of G. Hence a system G_r is optimal if $\|G - G_r\|_H = \sigma_{r+1}$.
It is important to note that the Hankel norm can is related to the L_∞ norm through the Nehari Theorem

$$G \in H_\infty, \quad F \in H_\infty^-, \quad G - F \in Ł_\infty \tag{2.82}$$

$$\|G\|_H = \min_{F \in H_\infty^-} \|G - F\|_\infty \tag{2.83}$$

.
Also the Adamjan-Arov-Krein theorem has to be stated to find a optimal solution to the stated minimization problem

$$G \in H_\infty, \quad Q \in H_\infty^-(r), \quad G - Q \in L_\infty \tag{2.84}$$

$$\min_{Q \in H_\infty^-} \|G - Q\|_\infty = \sigma_{r+1} \tag{2.85}$$

.
Suppose there is an optimal system $Q^* = G_r + F$ with $Q^* \in H_\infty^-(r), G_r \in H_\infty, F \in H_\infty^-$. It has the following error bound

$$\|G - G_r\|_\infty = \|G - Q^* + F\|_\infty \leq \sigma_{r+1} + \|F\|_\infty. \tag{2.86}$$

If $\|F\|_\infty$ is small enough, the stable part of Q^* can be used as a reduced order model. It is also an optimal solution to 2.80

$$\|G - G_r\|_H = \min_{F \in H_\infty^-} \|G - G_r - F\|_\infty = \min_{Q \in H_\infty^-} \|G - Q\|_\infty = \sigma_{r+1} \tag{2.87}$$

.
[32]

2.5.3. Constructing Q^*

The first step to construct an optimal system Q^* is to construct an balanced realization of the system $G = (A, B, C, D) \in H_\infty$that is to be reduced, as described in section 2.3.1. Hence the gramians W_c and W_o are equal and diagonal.

$$W_c = \begin{bmatrix} P_1 & 0 \\ 0 & \sigma_{r+1} I_l \end{bmatrix}, \quad W_o = \begin{bmatrix} Q_1 & 0 \\ 0 & \sigma_{r+1} I_l \end{bmatrix} \tag{2.88}$$

Now G is partitioned in the following way

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \tag{2.89}$$

.
Also a unitary matrix U has to be defined such tat $B_2 = -C_2^T U$ and $U^T U = I$. Further more a matrix $E_1 = P_1 Q_1 - \sigma_{r+1}^2 I$ is introduced.
Then a system $Q^* = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ can be defined

$$\hat{A} = E_1^{-1}(\sigma_{r+1}^2 A_{11}^T + Q_1 A_{11} P_1 - \sigma_{r+1} C_1^T U B_1^T) \tag{2.90}$$

$$\hat{B} = E_1^{-1}(Q_1 B_1 + \sigma_{r+1} C_1^T U) \tag{2.91}$$

$$\hat{C} = C_1 P_1 + \sigma_{r+1} U B_1^T \tag{2.92}$$

$$\hat{D} = D - \sigma_{r+1} U \tag{2.93}$$

. [32]

2.5.4. Decomposition of Q^*

The final step is to decompose Q^* into two systems $G_r \in H_\infty, F \in H_\infty^-$. To achieve this Q^* can be expressed in the diagonal canonical form described in section 2.54.

$$Q^* = D + \sum_{i=1}^{r} \frac{\phi_i}{s - \lambda_i} \tag{2.94}$$

.
This can no be decomposed into three systems $K \in H_\infty, V \in H_\infty^-$ and $\tilde{D}$ with

$$Re(\lambda_i) \in \mathbb{C}_{circ} \forall i \in I_\circ \tag{2.95}$$

$$K = \sum_{i \in I_-} \frac{\phi_i}{s - \lambda_i} \tag{2.96}$$

$$V = \sum_{i \in I_+} \frac{\phi_i}{s - \lambda_i} \tag{2.97}$$

$$\tilde{D} = \hat{D} \tag{2.98}$$

.
Since $\tilde{D}$ is some constant it does not have any poles. Therefore $K + D \in H_\infty$ and $V + D \in H_\infty^-$ which leads to two different decompositions $Q^* = (K + \tilde{D}) + V$ and $Q^* = K + (V + \tilde{D})$. From 2.86 it is clear that $\|F\|_\infty$ has to be as small as possible to minimize the error bound. It can be shown that $Q^* = (K + \tilde{D}) + V$ is the optimal decomposition. Suppose $Q^* = K + (V + \tilde{D})$

was the optimal decomposition, then the according lower bound has to be smaller

$$\|G - K\|_\infty \leq \|G - (K + \tilde{D})\|_\infty \tag{2.99}$$
$$\sigma_{r+1}\|V + \tilde{D}\|_\infty \leq \sigma_{r+1}\|V\|_\infty \tag{2.100}$$
$$\|V + \tilde{D}\|_\infty \leq \|V\|_\infty + \|\tilde{D}\|_\infty \geq \|V\|_\infty \tag{2.101}$$
$$\Rightarrow \|G - K\|_\infty \geq \|G - (K + \tilde{D})\|_\infty \tag{2.102}$$

.
Therefore $G_r = K + D$ and $F = V$.

2.6. Applying HNA to Heat Equation

To apply HNA to heat equation a state space representation of the system is needed to compute a reduced order system as described in the previous section. How the state space is constructed is already covered in section 2.3.4. However in contrast to the other mentioned methods, recovering an approximation of the full state solution is not trivial. Therfore in further analysis of the reduced system only error in frequency domain will be considered for HNA.

3. Implementation

The implementation of the discussed methods for solving the heat equation and for model order reduction was done in Matlab. Matlab was chosen as the programming language because it natively features matrix multiplication which finds heavy use in the previously mentioned methods. The second reason for this selection is that there exist ToolBoxes that already implement certain model order reduction methods such as MORLAB [33] or MOR toolbox [34]. The following figure shows the class diagram of the implementation:

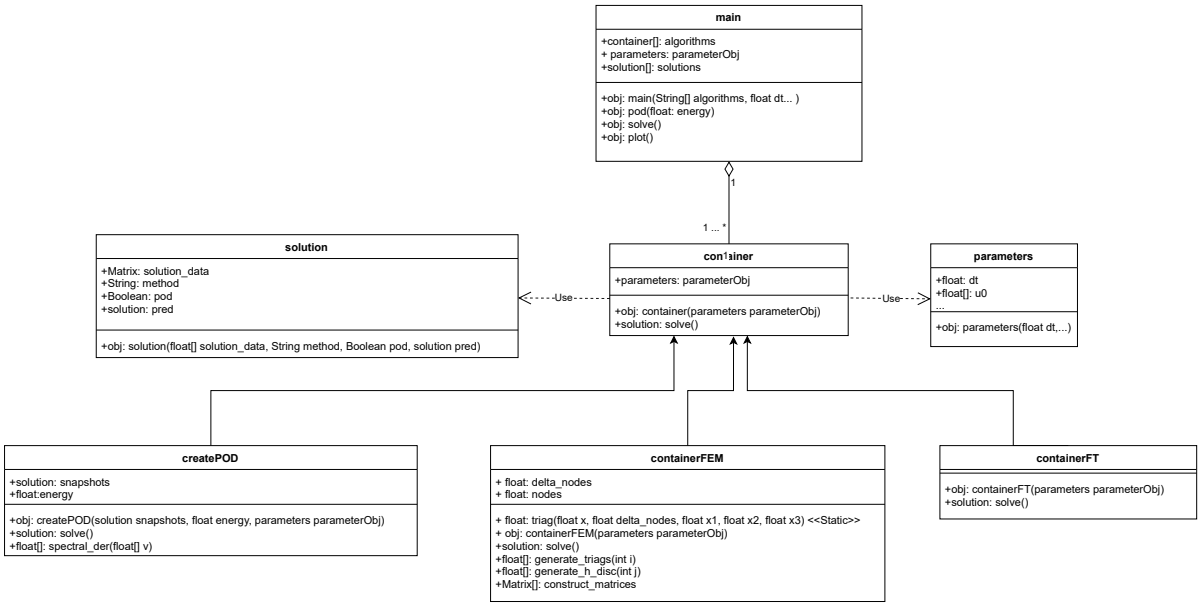

Figure 3.1.: Class diagramm of MOR and FEM implementation

3.1. Class main

The class main is responsible for generating the finite element solution, model order reduction steps and plotting the results. The process can be seen in the following flow chart:

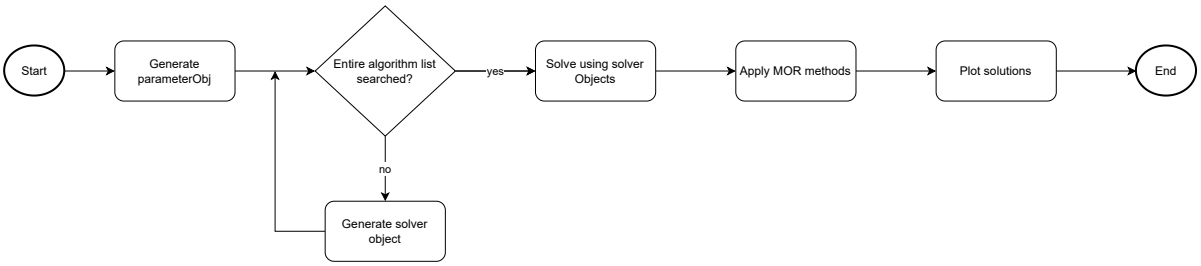

Figure 3.2.: Flow chart of main class

The first step is to generate a parameter object. The parameter object stores all parameters in order to increase the transparency and robustness of the program. The second step is to iterate the array of stated algorithms to solve the heat equation. The options are to solve the heat equation using finite element method 1.2 or using Fourier transform 1.1. After all solver objects have been generated, the according solutions are being computed. After that, the MOR methods are displayed. The final step is to display the solutions.

3.2. Class containerFEM

The class containerFEM is responsible for generating a solution using finite element method. FEM is implemented in the following way:

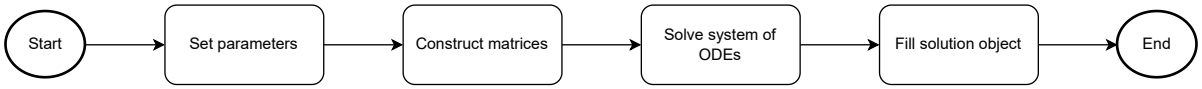

Figure 3.3.: Flow chart for FEM class

The first step is to set the parameters. After that the matrices discussed in 1.2 are being constructed. The next step is to solve the resulting system of ODEs and pass the solution to a solution object.

3.2.1. Construct Matrices

As defined in 1.34 the matrices K and M have to be constructed. Also to compute the initial condition given by u_0 F has to be known 1.36. This is done by the following method:

Algorithm 1 Construct matrices K, M and F

$ii \leftarrow \frac{2}{3}\Delta nodes$
$ij \leftarrow \frac{1}{6}\Delta nodes$
$F \leftarrow zeros(nodes, 1)$
$K \leftarrow zeros(nodes)$
$M \leftarrow zeros(nodes)$
for $i = 1$ **to** $nodes$ **do**
    $F(i) \leftarrow trapz(\phi_i \cdot u_0)$
end for
for $i = 1$ **to** $nodes$ **do**
    $K(i,i) \leftarrow \frac{-2}{\Delta nodes}$
    $M(i,i) \leftarrow ii$
    if $i - 1 > 1$ **then**
        $K(i, i-1) \leftarrow \frac{1}{\Delta nodes}$
        $M(i, i-1) \leftarrow ij$
    end if
    if $i + 1 < nodes + 1$ **then**
        $K(i, i+1) \leftarrow \frac{1}{\Delta nodes}$
        $M(i, i+1) \leftarrow ij$
    end if
end for
return$[F, K, M]$

3.2.2. Solve System of ODEs

The most important step in the process of generating a solution is to solve the system of ordinary differential equations that FEM yields:

Algorithm 2 Solve system of ODEs using euler scheme

1: $[F, K, M] \leftarrow construct\_matrices()$
2: $C \leftarrow zeros(nodes, n\_time\_steps)$
3: $c_0 \leftarrow M^{-1}F$
4: $C(:, 1) \leftarrow c_0$
5: $M(1, :) \leftarrow [0, \cdots, 0]$
6: $M(end, :) \leftarrow [0, \cdots, 0]$
7: $N \leftarrow M^{-1}K$
8: **for** $t = 2$ **to** n_time_steps **do**
9: $\quad d \leftarrow generate\_h\_disc(t)$
10: $\quad c_n \leftarrow \Delta t N c_0 + h(t) + c_0$
11: $\quad c_0 \leftarrow c_n$
12: $\quad C(:, t) \leftarrow c_n$
13: **end for**
14: $S \leftarrow []$
15: **for** $t = 1$ **to** n_time_steps **do**
16: $\quad c \leftarrow C(:, t)$
17: $\quad interpol \leftarrow interp1(linspace(0,\ L,\ nodes), c,\ X)$
18: $\quad S(:, t) \leftarrow interpol$
19: **end for**
20: **return** $solution(S,\ "FEM",\ 0,\ 0)$

In the first line the matrices F, K and M are retrieved. After that the initial vector of coefficients c_0 is computed using LS 1.37 in line three. The next two following lines force the boundary conditions as stated in **??**. In the lines 8 to 13 1.40 is implemented. In the last step the coefficients are interpolated in spatial direction to fit the given domain X and stored in a solution object.

4. Comparison of MOR Methods

The previously mentioned methods for model order reduction will be compared regarding time domain error, frequency domain error and computational speed. The time domain error will be obtained by compareing the FEM solution to a given approximation. To get insights into the frequency domain error, the error system of a reduced order model will be analysed. The computational speed will be determined by measuring the time it takes to generate a

ROM. Here it is assumed that the implementations provided by MORLAB are programmed in a sufficiently effective manner.

4.1. Time Domain Error

The time domain error will defined as $\epsilon = Y - \hat{Y}$ where Y and $\hat{Y}$ denote the matrices storing the output of the systems G and G_r. Since the output matrices are usually rather large, it is inpractical to use ϵ directly. Therefore $\|\epsilon\|_{max}$ and $\|\epsilon\|_F$ will be considered. The maximumsnorm will be used to show the magnitude of the worst occuring error. Since this is subsceptible to spikes the error is also meassured using the Frobeniusnorm to get a meassure of the error that respects all data points. Here two aspects are interesting. The first one is how the errors behaves as r gets larger and the second aspect is how the error evolves over time for some fixed r. To get data about the first aspect $\|\epsilon\|_{max}$ and $\|\epsilon\|_F$ will be meassured for increasing r. Data about the second aspect will be gathered by calculating the norm of the error between the column vectors of Y and $\hat{Y}$ for some fixed r, $\epsilon_t = \|Y_t - \hat{Y}_t\|_2$ where A_t denotes the t^{th} column vector of A. Since here vectors are compared the euclidean norm is used since it is compatible to the frobenius norm. The chebychev norm can also be used as vector norm.

Literaturverzeichnis

[1] Gustafsson, B. "Heat Conduction". In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 255–262. URL: https://doi.org/10.1007/978-3-642-19495-5_16.

[2] Papula, L. "Gewöhnliche Differentialgleichungen". In: *Mathematik für Ingenieure und Naturwissenschaftler Band 2: Ein Lehr- und Arbeitsbuch für das Grundstudium*. Wiesbaden: Springer Fachmedien Wiesbaden, 2015, pp. 343–541. URL: https://doi.org/10.1007/978-3-658-07790-7_4.

[3] Gustafsson, B. "Finite Element Methods". In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 173–192. URL: https://doi.org/10.1007/978-3-642-19495-5_11.

[4] Gustafsson, B. "Polynomial Expansions". In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 117–123. URL: https://doi.org/10.1007/978-3-642-19495-5_7.

[5] Gustafsson, B. "Least Square Problems". In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 125–133. URL: https://doi.org/10.1007/978-3-642-19495-5_8.

[6] Brunton, S. L./ Kutz, J. N. "Singular Value Decomposition (SVD)". In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 3–46.

[7] Brunton, S. L./ Kutz, J. N. "Fourier and Wavelet Transforms". In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 47–83.

[8] Szabo, F. E. "U". In: *The Linear Algebra Survival Guide*. Ed. by Szabo, F. E. Boston: Academic Press, 2015, pp. 385–392. URL: https://www.sciencedirect.com/science/article/pii/B978012409520550028X.

[9] Naoko, S. *Geometric Interpretation of the Correlation between Two Variables*. 2018. URL: https://medium.com/@ns2586/geometric-interpretation-of-the-correlation-between-two-variables-4011fb3ea18e (visited on 03/01/2023).

[10] Oliver, K. *Math 19b: Linear Algebra with Probability*. 2011. URL: https://people.math.harvard.edu/%5C~knill/teaching/math19b_2011/handouts/lecture12.pdf (visited on 03/01/2023).

[11] Victor, L. *IAML: Dimensionality Reduc6on*. 2011. URL: https://www.inf.ed.ac.uk/teaching/courses/iaml/2011/slides/pca.pdf (visited on 03/01/2023).

[12] Douglas, B. "The Fundamentals of Control Theory". In: Academia, 2019, pp. 8–15.

[13] Douglas, B. "The Fundamentals of Control Theory". In: Academia, 2019, pp. 42–47.

[14] Benner, P. et al. "1 Model order reduction: basic concepts and notation". In: *Volume 1 System- and Data-Driven Methods and Algorithms*. Ed. by Benner, P. et al. Berlin, Boston: De Gruyter, 2021, pp. 1–14. URL: https://doi.org/10.1515/9783110498967-001.

[15] Brunton, S. L./ Kutz, J. N. "Linear time-invariant systems". In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, p. 334.

[16] Douglas, B. "The Fundamentals of Control Theory". In: Academia, 2019, pp. 48–63.

[17] Abell, M. L./ Braselton, J. P. "Chapter 8 - Introduction to the Laplace Transform". In: *Introductory Differential Equations (Fifth Edition)*. Ed. by Abell, M. L./ Braselton, J. P. Fifth Edition. Academic Press, 2018, pp. 399–460. URL: https://www.sciencedirect.com/science/article/pii/B9780128149485000082.

[18] Brunton, S. L./ Kutz, J. N. "CONTROLLABILITY AND OBSERVABILITY". In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 335–341.

[19] *Volume 1 System- and Data-Driven Methods and Algorithms*. Berlin, Boston: De Gruyter, 2021. URL: https://doi.org/10.1515/9783110498967.

[20] Gustafsson, B. "Polynomial Expansions". In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 117–123. URL: https://doi.org/10.1007/978-3-642-19495-5_7.

[21] Gustafsson, B. “Numerical Methods for Differential Equations”. In: *Fundamentals of Scientific Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 137–144. URL: https://doi.org/10.1007/978-3-642-19495-5_9.

[22] Brunton, S. L./ Kutz, J. N. “Reduced Order Models (ROMs)”. In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 375–402.

[23] Brunton, S. L./ Kutz, J. N. “Discrete Fourier transform (DFT) and fast Fourier transform (FFT)”. In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 71–73.

[24] Ruppert, F. *Differentiation mit Matrizen*. matheplanet.com. URL:https://www.matheplanet.com (version: 2014-03-13). eprint: https://www.matheplanet.com/matheplanet/nuke/html/viewtopic.php?rd2\&topic=128338\&start=0\#p937673. URL: https://www.matheplanet.com/matheplanet/nuke/html/viewtopic.php?rd2%5C&topic=128338%5C&start=0%5C#p937673.

[25] Kloos, J. *When is matrix multiplication commutative?* Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/170371 (version: 2014-03-13). eprint: https://math.stackexchange.com/q/170371. URL: https://math.stackexchange.com/q/170371.

[26] Sapir, D. M. V. *Symmetric, diagonal, triangular matrices*. Linear Algebra Web-Notes. URL: https://math.vanderbilt.edu/sapirmv/msapir/jan22.html#:~:text=following%20result%20holds.-,Theorem.,B%20then%20AB%20is%20symmetric..

[27] Brunton, S. L./ Kutz, J. N. “BALANCED MODELS FOR CONTROL”. In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019, pp. 376–388.

[28] Vuillemin, P./ Maillard, A./ Poussot-Vassal, C. *Optimal Modal Truncation*. 2020. arXiv: 2009.11540 [math.OC].

[29] Peter Benner, L. F. *Model Reduction for Dynamical Systems — Lecture 4 —*. 2015. URL: https://www.mpi-magdeburg.mpg.de/2925234/lecture4.pdf (visited on 04/03/2023).

[30] Singh, N. “Reduction of linear dynamic systems using hankel norm approximation”. In: *Control and Intelligent Systems* 41 (01/2013).

[31] GLOVER, K. “All optimal Hankel-norm approximations of linear multivariable systems and their L, ∞-error bounds†”. In: *International Journal of Control* 39.6 (1984), pp. 1115–1193. eprint: https://doi.org/10.1080/00207178408933239. URL: https://doi.org/10.1080/00207178408933239.

[32] Sandberg, H. “Introduction to Model Order Reduction”. In: (07/2019), pp. 45–48. URL: http://www.diva-portal.org/smash/get/diva2:1338145/FULLTEXT01.pdf (visited on 04/15/2023).

[33] Benner, P./ Werner, S. W. R. *Morlab*. URL: https://www.mpi-magdeburg.mpg.de/projects/morlab (visited on 02/27/2023).

[34] Systems, P. D. *MOR toolbox*. URL: https://mordigitalsystems.fr/en/products.html (visited on 02/27/2023).

[35] D‘Erchie, M. e. a. “Analysis Eins”. In: *Mathe für Nicht-Freaks*. Serlo Education e. V., 2017, pp. 111–113. URL: https://mathe-builds.serlo.org/analysis1_final/analysis1_final.pdf.

[36] Alexandre M. Bayen, T. S. “An Introduction to MATLAB® Programming and Numerical Methods for Engineers”. In: Science Direct, 2015, pp. 211–214. URL: https://www.sciencedirect.com/book/9780124202283/an-introduction-to-matlab-programming-and-numerical-methods-for-engineers.

A. Appendix

A.1. Deriving matrices for FEM using piecewise linear functions

The so called triangle function is defined as follows:

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/\Delta x), & x_{j-1} \leq x < x_j \\ (x_{j+1} - x)/\Delta x), & x_j \leq x < x_{j+1} \\ 0, \quad \text{otherwise} \end{cases} \tag{A.1}$$

[4]
The following integrals have to be evaluated:

$$\int_\chi \phi_j \phi_k dx \quad \forall \phi_k \in \phi \tag{A.2}$$

$$-\int_\chi \frac{d\phi_j}{dx}\frac{d\phi_k}{dx} dx \quad \forall \phi_k \in \phi \tag{A.3}$$

With $\chi \subset \mathbb{R}$. Note that the product of two functions ϕ_j and ϕ_k and their derivatives is only under two conditions not zero:

1. $k = j$

Considering this case the integral A.2 becomes:

$$\int_{x_{j-1}}^{x_j} \phi_j^2 dx + \int_{x_j}^{x_{j+1}} \phi_j^2 dx \tag{A.4}$$

Because of symmetry only one of the above integrals have to computed:

$$2\int_{x_{j-1}}^{x_j}\phi_j^2 dx \tag{A.5}$$

$$=\frac{2}{\Delta x^2}\int_{x_{j-1}}^{x_j}(x-x_{j-1})^2 dx \tag{A.6}$$

$$\frac{2}{3\Delta x^2}\left[(x-x_{j-1})^3\right]_{x_{j-1}}^{x_j}=\frac{2}{3\Delta x^2}\Delta x^3=\frac{2}{3}\Delta x \tag{A.7}$$

Integral A.3 for $i=j$ taking symmetry into account becomes:

$$-\int_{x_{j-1}}^{x_{j+1}}(\frac{d\phi_j}{dx})^2 dx=-\frac{1}{\Delta x^2}\int_{x_{j-1}}^{x_{j+1}}1dx \tag{A.8}$$

$$=-\frac{1}{\Delta x^2}\left[x\right]_{x_{j-1}}^{x_{j+1}}=-\frac{2}{\Delta x} \tag{A.9}$$

2. $|j-k|=1$ A.2 becomes:

$$\frac{1}{\Delta x^2}\int_{x_j}^{x_{j+1}}(x-x_j)(x_{j+1}-x)dx \tag{A.10}$$

$$=\left[\frac{1}{2}x^2x_{j+1}-\frac{1}{3}x^3-xx_{j+1}x_j+\frac{1}{2}x^2x_j\right]_{x_j}^{x_{x_j+1}}=\frac{1}{6\Delta x^2}\Delta x^3=\frac{1}{6}\Delta x \tag{A.11}$$

Finally A.3 has to be evaluated for this condition:

$$-\int_{x_j}^{x_{j+1}}\frac{d\phi_j}{dx}\frac{d\phi_{j+1}}{dx}dx=\frac{1}{\Delta x^2}\int_{x_j}^{x_{j+1}}1dx=\frac{1}{\Delta x^2}\left[x\right]_{x_j}^{x_{j+1}}=\frac{1}{\Delta x} \tag{A.12}$$

A.2. Proof that matrix M is invertible

Let M_n be a matrix with $M_n \in \mathbb{R}^{n\times n}$ given by:

$$m_{ij} = \begin{cases} a, & k = j \\ b, & |k-j| = 1 \\ 0, & otherwise \end{cases} \tag{A.13}$$

It's determinant is given by the Laplace expansion:

$$det(M_n) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} N_{ij} \quad \forall i \tag{A.14}$$

N_{ij} is the determinant of the matrix M' that is obtained by removing the i^{th} row and j^{th} column of M_n. This expression can be simplified using the definition of m_{ij}:

$$det(M_n) = aN_{11} - bN_{12} \tag{A.15}$$

N_{11} is equivalent to $det(M_{n-1})$, since the indices of rows and columns of M' are in consecutive order and M' is a $n-1 \times n-1$ matrix :

$$N_{11} = det(M') \tag{A.16}$$

$$M' = \begin{bmatrix} m_{22} & \dots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n2} & \dots & m_{nn} \end{bmatrix} \tag{A.17}$$

N_{12} can be obtained by calculating the determinant of M' using the Laplace expansion:

$$M' = \begin{bmatrix} m_{21} & m_{23} & \dots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n3} & \dots & m_{nn} \end{bmatrix} \tag{A.18}$$

$$det(M') = m_{21} \cdot det(M'') \tag{A.19}$$

$$M'' = \begin{bmatrix} m_{33} & \dots & m_{3n} \\ \vdots & \ddots & \vdots \\ m_{n3} & \dots & m_{nn} \end{bmatrix} \tag{A.20}$$

In A.19 only the stated term has to be evaluated since all entries of the first column of the second sub matrix are zero. Therefore the determinant is zero. The row and column indices of M'' are in consecutive order and it is a $n-2 \times n-2$ matrix. Therefore M'' is equivalent to M_{n-2}. A.15 becomes:

$$det(M_n) = a \cdot det(M_{n-1}) - b^2 \cdot det(M_{n-2}) \tag{A.21}$$

Furthermore this implies $det(M_0) = 1$:

$$det(M_2) = a^2 - b^2 = a \cdot det(M_1) - b^2 \cdot 1 \tag{A.22}$$

$$\Rightarrow det(M_0) = 1 \tag{A.23}$$

Using the definition of 1.34 and 1.2.1 this can be seen as the following sequence:

$$a_0 = 1,\ a_1 = \frac{2\Delta x}{3} \tag{A.24}$$

$$a_{n+1} = \frac{2\Delta x}{3} \cdot a_n - \frac{\Delta x^2}{36} \cdot a_{n-1} \tag{A.25}$$

As described here [35] a recursive sequence converges if it is monotone and has a limit. A proof by induction shows that this sequence is monotone for $n \geq 1$.

Base case:

$$a_2 = (\frac{2\Delta x}{3})^2 - \frac{\Delta x^2}{36} = \Delta x^2(\frac{4}{9} - \frac{1}{36}) < \frac{2\Delta x}{3} = a_1 \tag{A.26}$$

Induction step: Assuming that $a_k < a_{k-1}$ holds, $a_{k+1} < a_k$ also holds:

$$a_{k+1} = \frac{2\Delta x}{3} \cdot a_k - \frac{\Delta x^2}{36} \cdot a_{k-1} < \frac{2\Delta x}{3} \cdot a_{k-1} - \frac{\Delta x^2}{36} \cdot a_{k-2} = a_k \tag{A.27}$$

The limit of this sequence is as follow:

$$\alpha = \lim_{n\to\infty} a_{n+1} = \lim_{n\to\infty} \Delta x \cdot \frac{2}{3} \cdot \lim_{n\to\infty} a_n - \lim_{n\to\infty} \Delta x^2 \cdot \frac{1}{36} \cdot \lim_{n\to\infty} a_{n-1} = 0 \cdot \alpha - 0 \cdot \alpha = 0 \tag{A.28}$$

Since this series is monotone and converges to zero as n goes to infinity, there is no $n \in \mathbb{N}$ for which $a_n = 0$. Therefore the determinant of the matrix M defined in 1.34 is not zero and M is invertible.

A.3. Equivalence of picewise linear polynomials and linear interpolation

A picewise linear polynomial in the form of:

$$u(x,t) = \sum_{j=1}^{N} c_j(t)\phi_j(x) \tag{A.29}$$

With ϕ_j being defined as A.1 and $c_j : \mathbb{R} \to \mathbb{R}$ is equivalent to linear interpolation with respect to x:

$$\hat{u}(x,t) = u_j + \frac{(u_{j+1} - u_j)(x - x_j)}{x_{j+1} - x_j} \tag{A.30}$$

for $x_j < x < x_{j+1}$ [36]. This can be shown by evaluating $u(x,t)$ between two neigbouring ϕ and $x_j < x < x_{j+1}$:

$$u(x,t) = \phi_j(x)c_j(t) + \phi_{j+1}(x)c_{j+1}(t) \tag{A.31}$$

$$= \frac{x_{j+1} - x}{\Delta x}c_j(t) + \frac{x - x_j}{\Delta x}c_{j+1}(t) \tag{A.32}$$

$$= \frac{x_{j+1} - x}{\Delta x}u_j + \frac{x - x_j}{\Delta x}u_{j+1} \tag{A.33}$$

$$= \frac{(x_{j+1}u_j - xu_j) + (xu_{j+1} - x_ju_{j+1})}{\Delta x} \tag{A.34}$$

$$= \frac{(u_{j+1} - u_j)x + x_{j+1}u_j - x_ju_{j+1}}{\Delta x} \tag{A.35}$$

$$= \frac{(u_{j+1} - u_j)x + x_ju_j + \Delta xu_j - x_ju_{j+1}}{\Delta x} \tag{A.36}$$

$$= u_j + \frac{(u_{j+1} - u_j)x - x_j(u_{j+1} - u_j)}{\Delta x} \tag{A.37}$$

$$= u_j + \frac{(u_{j+1} - u_j)(x - x_j)}{x_{j+1} - x_j} \tag{A.38}$$